Uniting genetic and geographic databases to understand the relationship between latitude and population demography

Frank Burbrink[1]

[1]American Museum of Natural History

June 14, 2022

Abstract

Conducting large-scale phylogeographic studies to understand processes affecting population structure and genetic diversity across multiple species is difficult because the key genetic (NCBI) and spatial (GBIF) repositories are disconnected. In this issue of Molecular Ecology Resources, Pelletier et al. (2022) demonstrate the power of connecting these in the program phylogatR. This program assembled 87,852 species and 102,268 sequence alignments in a taxonomic hierarchy, yielding multiple sequence alignments per species, mainly for animals (88%), composed mostly of mtDNA data. The authors discuss several caveats with these alignments and provide flags identifying particular problems associating locality and genetic data with certain taxa (e.g., multiple localities per individuals). They provide a test that nucleotide diversity should increase with area, but find a significant relationship in only 32% of taxa with no clear taxonomic or ecological factors accounting for this. To examine the potential of this program, I tested the idea that the degree of population expansion should increase with latitude given potential environmental stability in the tropics and instability in temperate regions. In under two hours, I downloaded all squamates (lizards and snakes) and regressed Tajima's D on latitude and found a weak but significant negative relationship, indicating a potential association between latitude and population expansion. The phylogatR database is a powerful resource for researchers wanting to test the relationship between genetic diversity and some aspect of space or environment.

Title: Uniting genetic and geographic databases to understand the relationship between latitude and population demography

Running Title: Aggregating geography and genetic databases

Frank T. Burbrink

Department of Herpetology

American Museum of Natural History

Central Park West at 79th Street

New York, NY 10024-5192

fburbrink@amnh.org

Keywords: *phylogatR,* phylogeography, Tajima's D, latitude

Abstract:

Conducting large-scale phylogeographic studies to understand processes affecting population structure and genetic diversity across multiple species is difficult because the key genetic (NCBI) and spatial (GBIF) repositories are disconnected. In this issue of *Molecular Ecology Resources* , Pelletier et al. (2022) demonstrate the power of connecting these in the program *phylogatR.* This program assembled 87,852 species and 102,268 sequence alignments in a taxonomic hierarchy, yielding multiple sequence alignments per species, mainly for

animals (88%), composed mostly of mtDNA data. The authors discuss several caveats with these alignments and provide flags identifying particular problems associating locality and genetic data with certain taxa (e.g., multiple localities per individuals). They provide a test that nucleotide diversity should increase with area, but find a significant relationship in only 32% of taxa with no clear taxonomic or ecological factors accounting for this. To examine the potential of this program, I tested the idea that the degree of population expansion should increase with latitude given potential environmental stability in the tropics and instability in temperate regions. In under two hours, I downloaded all squamates (lizards and snakes) and regressed Tajima's D on latitude and found a weak but significant negative relationship, indicating a potential association between latitude and population expansion. The *phylogatR* database is a powerful resource for researchers wanting to test the relationship between genetic diversity and some aspect of space or environment.

It was not known that phylogenetic analysis of population genetic data would show geographic structure when Avise et al. (1979) introduced the field of phylogeography, so named eight years later (Avise et al., 1987). Since then, the field has grown from using single gene fragments to whole genomes resulting in more than 22,000 publications. This seemingly simple relationship between geography and genetic variation has provided the foundation for studying speciation, species delimitation, hybrid zone dynamics, adaptation, conservation genetics, community assembly, historical demography, and climate change response to name a few (Frank T. Burbrink et al., 2016; F. T. Burbrink & Ruane, 2021; Carnaval et al., 2009; Dapporto et al., 2009; Dufresnes et al., 2020; Hewitt, 2001; Overcast et al., 2019; Rissler & Smith, 2010; Satler & Carstens, 2017; Shaffer et al., 2004; Smith et al., 2011; Soltis et al., 2006).

As with many burgeoning fields, there is often little consideration of how to make datasets accessible for future researchers addressing more comprehensive questions under a common framework. For example, it is common to examine how shared environments or barriers affect population structure across communities of species, or test if range size or latitude are correlated with genetic diversity across taxa (Hickerson et al., 2006; Myers et al., 2019; Smith, et al., 2017). However, addressing these types of questions using existing data requires researchers to assemble large databases manually. Genetic and geographic databases used to store this information like NCBI Genbank (National Center for Biotechnology Information) and GBIF (Global Biodiversity Information Facility) are disconnected, and often of limited general use for conducting multitaxon studies. In this issue of *Molecular Ecology* , Pelletier et al. (2022) have automated the process of connecting geography to DNA sequences via the *phylogatR* (phylogeographic data aggregation and repurposing) database.

The *phylogatR* database has assembled 87,852 species and 102,268 sequence alignment and associated spatial data. The database represents mostly animals (88%) distantly followed by plants (9%) generated from NCBI Genbank, BOLD (Barcode of Life Database), and GBIF. This program is automatically updated monthly for new entries. The alignments produced by *phylogatR* are generated by MAFFT v7 (Katoh & Standley, 2013), checked for alignment and gap issues, and are ready to use for analyses. To note, the authors have developed a system for flagging sequences with potential problems, such as multiple unique geographic coordinates referenced to a single sequenced sample or changes in taxonomy. Pelletier et al. (2022) have provided several tutorials that explain how to use the database, which should be useful for teachers conducting workshops

The database now has 2.6 million records representing 1988 genes, with most species having only 1.2 genes and 25.8 sequences per alignment. As expected, most of the sequence alignments here are represented by mitochondrial and chloroplast DNA. To provide a test of the data collected in *phylogatR,* Pelletier et al. (2022) ask if range size predicts nucleotide diversity (π), an old but important question in population genetics (Wright, 1943), but now using 80,000 species and over 2 million sequences. Nucleotide diversity was estimated in *Pegas*(Paradis, 2010) and was regressed against geographic area calculated from the associated georeferenced data. The authors discovered only 58 geographic outliers for taxa with large ranges and 23 π outliers, mostly due to mixed-gene alignments or individuals missing overlapping sequences. Interestingly, a majority of groups (68%) showed no significant relationship between area and nucleotide diversity. There seemed to be no taxonomic or general ecological trend among those groups showing a significant relationship. Of course many other factors might contribute to genetic diversity and this bears further exploration as the

authors suggest.

To test drive *phylogatR* , I conducted a study on squamates (lizards and snakes) to address a key question in evolutionary ecology: are tropical regions more stable than temperate regions and does this affect biodiversity (Dobzhansky, 1950)? Stability of taxa in the tropics relative to those in temperate regions with greater environmental fluctuation over time should show evidence of greater demographic expansions at higher latitudes (Lessa et al., 2003; Whorley et al., 2004). I estimated Tajima's D for each species and regressed these against latitude using R (R Core Team 2020). Negative values of D suggest population expansion from a bottleneck or a selective sweep, whereas values close to zero indicate neutrality (Stajich & Hahn, 2005; Tajima, 1989). I also examined the relationship between area and π. Because a majority of species only have mtDNA, I kept the longest fragment of mtDNA with the most individuals per species. This yielded a dataset of 418 species with an average of 13.59 individuals/species. I found a significant relationship between Tajima'sD and latitude ($P = 0.01$), though the effect size was small ($r^2 = 0.012$; Fig.1). Because low sampling can affect correct estimation of Tajima's D , I filtered the dataset to only include taxa with > 10 individuals (n = 143); this also generated a significant relationship with a weak effect ($P = 0.037$; $r^2 = 0.023$). The prediction that populations may be more stable in the tropics and that the magnitude of population expansion increases with latitude holds. However, the effect is weak; negative D might also be associated with selective sweeps, and relying on a single locus may not provide the strongest inference of population-level processes (Burbrink & Ruane, 2021). Interestingly, the mean value of D across all squamates was close to zero (-0.67), suggesting a strong role for neutrality. Similar to Pelletier et al. (2022), I found no relationship between area and π ($P = 0.19 – 0.36$), though the geographic extent here for some taxa may be inaccurate. This study was completed in ˜2 hours on a MacBook Pro.

Because phylogatR is a big-data aggregator, it is likely that more fine-scale problems with individual species alignments and georeferenced data are present. In the squamate study, I used an outlier detector (Grubbs test) for π and D and found two species (0.4%) with either individuals missing 95% of their data or lack overlap among most gene fragments. These kinds of problems could be identified prior to analyses with simple scripts that detect missing data beyond some user input threshold or sequence mismatch. I found geographic outliers were caused by taxa that had been introduced well beyond their natural range (e.g.., *Hemidactylus* geckos). This requires the end user to know about the natural history of their target study organisms and assess if this is a problem for their particular study design.

The program *phylogatR* represents a major leap forward for aggregating all of those phylogeographic datasets accumulating since Avise et al. (1979). The database is easy to use, is a major time saver, and the caveats are clear. I envision some future version of this that scrapes genome-scale data now also accumulating at a massive rate. In the meantime, the current version can facilitate the next generation of comparative ecological and community level analyses of phylogeographic patterns and processes.

Acknowledgements:

I thank A. Pyron for providing comments on the original draft. This perspective was written with support from the NSF (Dimensions-USBiota 1831241).

Data Accessibility Statement: All data are available in

phylogatR.

Benefits Generated: Benefits from this research accrue from the sharing data and results on public databases as described above

Author Contribution: FTB did everything for this paper.

References

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., Saunders, N. C. (1987). Intraspecific phylogeography - the mitochondrial-DNA bridge between population-genetics and systematics. *Annual Review of Ecology and Systematics*, *18*, 489–522.

Avise, J. C., Giblin-Davidson, C., Laerm, J., Patton, J. C., & Lansman, R. A. (1979). Mitochondrial DNA clones and matriarchal phylogeny within and among geographic populations of the pocket gopher, *Geomys pinetis*. *Proceedings of the National Academy of Sciences of the United States of America*,*76*(12), 6694–6698.

Burbrink, F. T., Chan, Y. L., Myers, E. A., Ruane, S., Smith, B. T., & Hickerson, M. J. (2016). Asynchronous demographic responses to Pleistocene climate change in Eastern Nearctic vertebrates. *Ecology Letters*, *19*(12), 1457–1467.

Burbrink, F. T., & Ruane, S. (2021). Contemporary philosophy and methods for studying speciation and delimiting species. *Ichthyology and Herpetology*, *109*(3), 874–894.

Carnaval, A. C., Hickerson, M. J., Haddad, C. F. B., Rodrigues, M. T., & Moritz, C. (2009). Stability predicts genetic diversity in the Brazilian Atlantic forest hotspot.*Science*, *323*(5915), 785–789.

Dapporto, L., Bruschini, C., Baracchi, D., Cini, A., Gayubo, S. F., González, J. A., & Dennis, R. L. H. (2009). Phylogeography and counter-intuitive inferences in island biogeography: evidence from morphometric markers in the mobile butterfly*Maniola jurtina* (Linnaeus) (Lepidoptera, Nymphalidae).*Biological Journal of the Linnean Society*, Vol. 98, pp. 677–692. doi: 10.1111/j.1095-8312.2009.01311.x

Dobzhansky, T. (1950). Evolution in the tropics. *American Scientist*, *38*(2), 208–221.

Dufresnes, C., Berroneau, M., Dubey, S., Litvinchuk, S. N., & Perrin, N. (2020). The effect of phylogeographic history on species boundaries: a comparative framework in *Hyla* tree frogs. *Scientific Reports*, *10*(1), 5502.

Hewitt, G. M. (2001). Speciation, hybrid zones and phylogeography - or seeing genes in space and time. *Molecular Ecology*, *10*(3), 537–549.

Hickerson, M. J., Dolman, G., & Moritz, C. (2006). Comparative phylogeographic summary statistics for testing simultaneous vicariance. *Molecular Ecology*, *15*(1), 209–223.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780.

Lessa, E. P., Cook, J. A., & Patton, J. L. (2003). Genetic footprints of demographic expansion in North America, but not Amazonia, during the Late Quaternary.*Proceedings of the National Academy of Sciences of the United States of America*, *100*(18), 10331–10334.

Myers, E. A., Xue, A. T., Gehara, M., Cox, C. L., Davis Rabosky, A. R., Lemos-Espinal, J., Burbrink, F. T. (2019). Environmental heterogeneity and not vicariant biogeographic barriers generate community-wide population structure in desert-adapted snakes. *Molecular Ecology*, *28*(20), 4535–4548.

Overcast, I., Emerson, B. C., & Hickerson, M. J. (2019). An integrated model of population genetics and community ecology. *Journal of Biogeography*, *46*(4), 816–829.

Paradis, E. (2010). pegas: an R package for population genetics with an integrated-modular approach.*Bioinformatics* , *26*(3), 419–420.

Pelletier, T., Parsons, D., Decker, S., Crouch, S., Franz, E., Ohrstrom, J., & Carstens, B. (2022). phylogatR: Phylogeographic data aggregation and repurposing.*Molecular Ecology Resources*, *In Press*.

R Core Team (2020). A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Rissler, L. J., & Smith, W. H. (2010). Mapping amphibian contact zones and phylogeographical break hotspots across the United States. *Molecular Ecology*,*19*(24), 5404–5416.

Satler, J. D., & Carstens, B. C. (2017). Do ecological communities disperse across biogeographic barriers as a unit? *Molecular Ecology*, *26*(13), 3533–3545.

Shaffer, H. B., Pauly, G. B., Oliver, J. C., & Trenham, P. C. (2004). The molecular phylogenetics of endangerment: cryptic variation and historical phylogeography of the California tiger salamander, *Ambystoma californiense.Molecular Ecology*, *13*(10), 3033–3049.

Smith, B. T., Escalante, P., Hernández Baños, B. E., Navarro-Sigüenza, A. G., Rohwer, S., & Klicka, J. (2011). The role of historical and contemporary processes on phylogeographic structure and genetic diversity in the Northern Cardinal, Cardinalis cardinalis. *BMC Evolutionary Biology*,*11*(1), 136.

Smith, B. T., Seeholzer, G. F., Harvey, M. G., Cuervo, A. M., & Brumfield, R. T. (2017). A latitudinal phylogeographic diversity gradient in birds. *PLoS Biology*, *15*(4), e2001073.

Soltis, D. E., Morris, A. B., McLachlan, J. S., Manos, P. S., & Soltis, P. S. (2006). Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology*, *15*(14), 4261–4293.

Stajich, J. E., & Hahn, M. W. (2005). Disentangling the effects of demography and selection in human history. *Molecular Biology and Evolution*, *22*(1), 63–73.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.*Genetics*, *123*(3), 585–595.

Whorley, J. R., Alvarez-Castañeda, S. T., & Kenagy, G. J. (2004). Genetic structure of desert ground squirrels over a 20-degree-latitude transect from Oregon through the Baja California peninsula. *Molecular Ecology*,*13*(9), 2709–2720.

Wright, S. (1943). Isolation by Distance. *Genetics*, *28*(2), 114–138.

Figure Legends

Figure 1. The top graph shows the relationship between Tajima's *D*regressed on latitude across squamates. The bottom graph shows the distribution of Tajima's *D* among squamates. For both graphs, the pink color represents the full dataset of 418 taxa and the blue color represents the reduced dataset of 143 taxa, filtering out species with less than 10 individuals.